In Silico Human and Rat V_{ss} Quantitative Structure−Activity Relationship Models

M. Paul Gleeson,*,† Nigel J. Waters,* Stuart W. Paine, and Andrew M. Davis

Department of Physical & Metabolic Sciences, AstraZeneca R&D Charnwood, Bakewell Road, Loughborough, Leicestershire LE11 5RH, United Kingdom

Received October 7, 2005

We present herein a QSAR tool enabling an entirely in silico prediction of human and rat steady-state volume of distribution (V_{ss}), to be made prior to chemical synthesis, preceding detailed allometric or mechanistic assessment of V_{ss}. Three different statistical methodologies, Bayesian neural networks (BNN), classification and regression trees (CART), and partial least squares (PLS) were employed to model human ($N = 199$) and rat ($N = 2086$) data sets. The results in prediction of an independent test set show the human model has an r^2 of 0.60 and an rms error in prediction of 0.48. The corresponding rat model has an r^2 of 0.53 and an rms error in prediction of 0.37, indicating both models could be very useful in the early stages of the drug discovery process. This is the first reported entirely in silico approach to the prediction of rat and human steady-state volume of distribution.

Introduction

The huge cost of pharmaceutical drug development (the current cost of discovering a new therapy is thought to approach U.S. $1.3−1.6 billion[1]) and the high attrition of compounds entering clinical development are rightly focusing attention upon every aspect of the efficiency of our industry. While reasons for attrition are varied, including portfolio decisions and lack of clinical efficacy of the biological mechanism, many reasons for compound failure are entirely controlled by the chemical structure. Therefore, there is still much that can be done in the discovery phase, to improve the chances of success of a candidate drug later in development, by the judicious choice of a chemical target. Hence, the current focus is upon the use of predictive ADMET[a] models allowing the biological properties of virtual structures to be predicted and a more informed choice of target to be selected for synthesis.

The prediction of the steady-state volume of distribution (V_{ss}) is a key pharmacokinetic parameter, which together with clearance determines the half-life, and thus impacts on the dosing regimen of a compound. The dosing regimen is designed to maintain a free plasma concentration, greater than that required to give the pharmacodynamic effect throughout the dosing interval, while lessening the maximal concentration (C_{max}) and potential for related side effects. These pharmacodynamic parameters, being very difficult to predict and unique to the pharmacological target, have meant that efforts have been concentrated on predicting half-life.[2] While allometric scaling[3] and correlative methods have been used, in vivo and in vitro data in animals are required and improved correlations tend to be achieved by prediction of the two major determinants, clearance and V_{ss}.[4]

V_{ss} represents the volume in which a drug would appear to be distributed during steady state if the drug existed throughout that volume at the same concentration as that in the measured fluid (blood or plasma). It is a function of binding to plasma and tissue components and as such is commonly expressed via the Gillette equation:[5]

$$V_{ss} = V_p + V_t\left(\frac{f_{up}}{f_{ut}}\right) \quad (1)$$

where V_p is the plasma volume, V_t is the tissue volume, and f_u is the fraction unbound in plasma (p) and tissue (t). While it is possible to measure f_{up} in man, it is not practical to measure f_{ut}, as f_{ut} represents a weighted mean for all tissues. While f_{ut} for individual tissues can be measured using dialysis, it becomes nonpractical for multiple tissues for rapid progression of compounds. Furthermore, excellent correlations between lipophilicity and unbound fraction in rat tissue, as measured by equilibrium dialysis, have been demonstrated.[6] The binding and partition phenomena, which drive the free fraction in plasma and tissue, are often explained in terms of physicochemical descriptors and thus indicate it should be possible to predict the V_{ss} of a compound purely from its structure.

A number of different strategies have been applied to the prediction of human V_{ss} including allometry, PBPK modeling,[7] and multivariate analysis of animal V_{ss} data.[8] These approaches, all requiring experimental data sets, have proven successful at the latter stages of lead optimization and in preclinical development. The application of quantitative structure activity relationship (QSAR) methods to predict human pharmacokinetics is a growing field, with the potential to reduce research and development time, costs, and resources. Recent examples include work by Lombardo et al.[9] and Ghafourian and colleagues[10] who used measured and calculated descriptors to predict V_{ss}. However, both investigations relied on relatively small data sets (∼100) and the requirement for measured data by these models means that they cannot be applied to virtual compounds.

In our study, models for human V_{ss} were constructed using purely calculated descriptors for a data set of 199 marketed drugs covering a range of molecular properties and V_{ss} values (Figure 1). This was conducted in parallel with model building for rat V_{ss} based on a data set of 2086 in-house compounds covering

* To whom correspondence should be addressed. Phone: +44-(0)1438-768682. Fax: +44-(0)1438-763352. E-mail: paul.x.gleeson@gsk.com, nigel.waters@astrazeneca.com.

† Current Address: GlaxoSmithKline Medicines Research Centre, Computational & Structural Sciences, Gunnels Wood Road, Stevenage, Hertfordshire, SG1 2NY, United Kingdom.

[a] Abbreviations: ADMET, adsorption, distribution, metabolism, excretion, and toxicity; BNN, Bayesian neural network; CART, classification and regression trees; LO, lead optimisation; ME, mean error; PLS, partial least squares; PPB, plasma protein binding; QSAR, quantitative structure−activity relationship; QSPR, quantitative structure−property relationship; rmse, root-mean-square error; VDW, van der Waals; VIP, variable importance; V_{ss}, steady-state volume of distribution.

10.1021/jm0510070 CCC: $33.50
Published on Web 02/18/2006

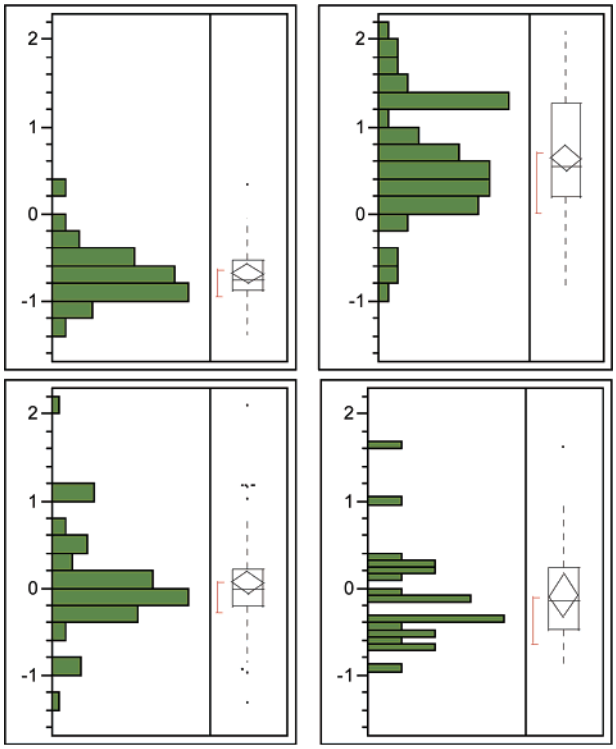

Figure 1. Experimental human V_{ss} distribution (log V_{ss}) for 199 compounds by charge type. The charge types are displayed clockwise from top left: acid ($N = 33$, mean $= -0.70$, standard deviation (SD) $= 0.32$), base ($N = 74$, mean $= 0.62$, SD $= 0.65$), zwitterion ($N = 23$, mean $= -0.10$, SD $= 0.57$), and neutral ($N = 69$, mean $= 0.06$, SD $= 0.56$).

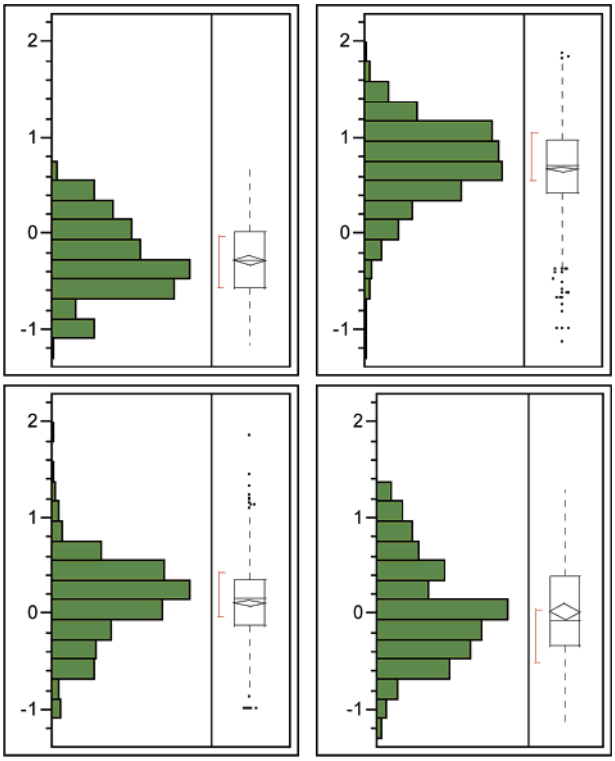

Figure 2. Experimental rat V_{ss} distribution (log V_{ss}) for 2086 compounds by charge type. The charge types are displayed clockwise from top left: acid ($N = 199$, mean $= -0.29$, SD $= 0.39$), base ($N = 994$, mean $= 0.66$, SD $= 0.43$), zwitterion ($N = 130$, mean $= 0.01$, SD $= 0.51$), and neutral ($N = 763$, mean $= 0.10$, SD $= 0.41$).

diverse chemical space and V_{ss} values (Figure 2). By employing three distinct but complementary QSAR methods—partial least squares (PLS), classification and regression trees (CART), and Bayesian neural networks (BNN)—we aimed to model V_{ss} using solely physicochemical descriptors generated in silico. The use of three distinctly different statistical methodologies also allows us to explore more completely the multidimensional molecular hyperplane that controls V_{ss} as we can explicitly account for both linear and nonlinear dependencies. We also investigate the effect of consensus predictions which have often proved more effective than predictions from individual models alone.

Results and Discussion

To compare the performances of our different QSAR models, we employ four different statistics: r_0^2 is the coefficient of determination, to the line of unity $y = x$; r^2 is the square of the Pearson's correlation coefficient, based on the line of best fit; rmse is the root-mean-square error in prediction; ME is the mean error in prediction. These statistics collectively allow us to determine the quality of the correlation, either in absolute terms (r_0^2) or in the rank ordering (r^2), the model error in the units of the measurement (rmse), and the presence or absence of any systematic bias (ME) in prediction.

1. Human QSAR Models. The results of the human V_{ss} models in fit and prediction are given in Table 2 and Table 3, respectively. The results in Table 2 show that we are describing between 64 and 87% of the total variance in the training set depending on the model; however, whether these statistics are reproduced in the independent test set cannot be guaranteed. Analysis of the individual training set fit results would suggest that the CART and BNN models will predict human V_{ss} more effectively than PLS on account of their larger r^2 and smaller rmse values. However, we often find the quite rigorous cross-validation procedures in PLS (leave $^1/_7$ of data out 7 times with reassessment of the model) ensure the model is not overfitted with respect to the training set, which often leads to similar performance in training set fit and test set prediction. In contrast, there can frequently be an increase in rmse in the CART and BNN models when comparing the training set fit to the test set predictions. This is clearly shown by comparison of the training set and test set model rmse values for each modeling approach in Tables 2 and 3. The BNN and CART models show test set statistics that deteriorate quite significantly from the training to test sets with the r^2 falling from ~0.79 to ~0.56. The PLS model is the most predictive on the test set, explaining 58% of the total variance, with the r^2 from the training to test showing a more modest decrease (0.64 to 0.58) compared to the CART or BNN models. The rmse prediction for all three models is ~0.5 log units, as compared to an experimental error of ~0.2 log units based on an analysis of the replicate measurements.

The predictive ability of the model can be improved further by taking a consensus of all three predictions. This results in the total variance of the test set being explained by the model increasing to 0.60 and the rmse decreasing to 0.48 (Figure 3). These in silico model results are in accord with models dependent on measured properties reported by others in the literature[9,10] but are made from an entirely computational procedure.

2. Rat QSAR Models. The rat training set results given in Table 4 might again suggest that the CART and BNN models would predict more effectively than the PLS on account of their larger r^2 values and lower rmse values in fit. Examination of the results in prediction of the independent test set of 416 compounds (Table 5) does show that the BNN model is the most predictive of the three followed by the PLS model and

Table 1. Literature Human V_{ss} Data Set (N =199)[a]

compd	V_{ss} (L/kg)	set	charge type	ref	compd	V_{ss} (L/kg)	set	charge type	ref
acebutolol	1.2	training	base	1	glyburide	0.2	training	acid	2
acyclovir	0.69	training	neutral	2	granisetron hydrochloride	3	training	base	2
alprazolam**	0.72	training	neutral	2					
amiloride	17	training	base	2	hydralazine	1.5	training	neutral	2
amiodarone	66	training	base	2	hydrochlorothiazide	0.83	training	neutral	2
amitriptyline	15	training	base	2	ibuprofen	0.15	training	acid	2
amoxicillin	0.21	training	zwitterion	2	imipramine	18	training	base	2
ampicillin	0.28	training	zwitterion	2	indomethacin	0.29	training	acid	2
atropine	2	training	base	2	isradipine	4	training	neutral	2
auranofin**	0.045	training	neutral	2	itraconazole	14	training	neutral	2
azithromycin	31	training	base	2	ketoconazole	0.15	training	base	2
bepridil	8	training	base	2	ketoprofen	0.15	training	acid	2
bisoprolol	3.2	training	base	2	ketorolac	0.21	training	acid	2
bromocriptine mesylate	2	training	base	2	lomefloxacin	2.3	training	zwitterion	2
					loracarbef	0.32	training	zwitterion	2
bupropion	7.2	training	base	2	loratidine	120	training	neutral	2
caffeine	0.61	training	neutral	1	lorazepam	1.3	training	neutral	2
captopril	0.81	training	acid	2	maprotiline	43	training	base	1
carbamazepine	1.4	training	neutral	2	mefloquine	19	training	base	2
carbenicillin	0.18	training	acid	2	meperidine	4.4	training	base	2
cefaclor	0.36	training	zwitterion	2	methadone	3.8	training	base	2
cefprozil	0.22	training	zwitterion	2	methyldopa	0.46	training	zwitterion	2
cephalexin	0.26	training	zwitterion	2	methylprednisolone	1.2	training	neutral	2
chlorambucil	0.29	training	acid	2	metronidazole	0.74	training	neutral	2
chloramphenicol	0.94	training	neutral	2	mexiletine	4.9	training	base	2
chlorothiazide	0.2	training	base	2	minocycline	1.3	training	zwitterion	2
chlorpheniramine maleate	3.2	training	base	2	morphine sulfate	3.3	training	base	2
					nabumetone	0.79	training	neutral	2
chlorpropamide	0.097	training	acid	2	nafcillin sodium	0.35	training	acid	2
chlorthalidone	0.1	training	neutral	2	naloxone	2.1	training	base	2
cimetidine	1	training	base	2	naltrexone hydrochloride	19	training	base	2
cinoxacin	0.33	training	acid	2					
ciprofloxacin	1.8	training	zwitterion	2	naproxen sodium	0.16	training	acid	2
clarithromycin	2.6	training	base	2	nicardipine	1.1	training	base	2
clavulanate	0.21	training	acid	2	nifedipine	0.78	training	neutral	2
clindamycin	1.1	training	base	2	nimodipine	1.7	training	neutral	2
clofibrate	0.11	training	neutral	2	nitrofurantoin	0.58	training	neutral	2
clomipramine	20	training	base	1	nizatidine	1.2	training	base	2
cyclophosphamide	0.78	training	neutral	2	norethindrone	3.6	training	neutral	2
cyclosporine	1.3	training	neutral	2	norfloxacin	0.7	training	zwitterion	2
dexamethasone	0.82	training	neutral	2	nortriptyline	18	training	base	2
diazepam	1.3	training	neutral	1	ofloxacin	1.8	training	zwitterion	2
diazoxide	0.21	training	base	2	omeprazole	0.34	training	neutral	2
dicloxacillin	0.086	training	acid	2	oxaprozin	0.19	training	acid	2
didanosine	1	training	neutral	2	oxazepam	0.6	training	neutral	2
diflunisal	0.1	training	acid	2	paroxetine	0.26	training	base	2
digoxin	3.12	training	neutral	2	pentazocine	7.1	training	base	2
dihydrocodeine	3.1	training	base	2	phenobarbital	0.54	training	neutral	2
doxazosin	1.5	training	neutral	2	phenytoin	0.64	training	neutral	2
doxepin	20	training	base	2	pimozide	28	training	base	2
enalapril	1.7	training	zwitterion	2	pindolol	2.3	training	base	2
erythromycin	0.78	training	base	2	piroxicam	0.15	training	neutral	2
ethambutol**	1.6	training	base	2	prazepam	14.4	training	neutral	2
etodolac	0.36	training	acid	2	prazosin	0.6	training	neutral	2
famotidine	1.3	training	base	2	prednisolone	1.5	training	neutral	2
felbamate	0.76	training	neutral	2	primidone	0.69	training	neutral	2
felodipine	10	training	neutral	2	procainamide	1.9	training	base	2
finasteride	1.1	training	neutral	2	propranolol	4.3	training	base	2
fluconazole	0.6	training	neutral	2	pyrazinamide	0.7	training	neutral	2
flucytosine	0.68	training	neutral	2	pyrimethamine	2.3	training	neutral	2
fluoxetine	35	training	base	2	quinapril	0.4	training	zwitterion	2
flurazepam hydrochloride	22	training	base	2	quinine sulfate	1.8	training	base	2
					ranitidine	1.3	training	base	2
flurbiprofen	0.15	training	acid	2	rifabutin	40	training	zwitterion	2
fosinopril sodium	0.13	training	acid	2	rifampin	0.97	training	zwitterion	2
furosemide	0.11	training	acid	2	rimantadine hydrochloride**	25	training	base	2
ganciclovir	1.1	training	neutral	2					
gemfibrozil	0.14	training	acid	2	risperidone	1.1	training	base	2
glipizide	0.17	training	acid	2	spironolactone	14	training	neutral	2
sulfamethoxazole	0.29	training	acid	2	diphenhydramine	4.5	test	base	2
sulfinpyrazone	0.74	training	neutral	2	doxycycline	0.75	test	zwitterion	2
sumatriptan succinate	0.65	training	base	2	ethinyl estradiol	3.5	test	neutral	2
tacrine hydrochloride	5.9	training	neutral	2	etoposide	0.36	test	neutral	2
tacrolimus	0.88	training	neutral	2	famciclovir	0.98	test	neutral	2
tamoxifen citrate	55	training	base	2	flecainide	4.9	test	base	2
temazepam	0.95	training	neutral	2	haloperidol	18	test	base	2

Table 1 (Continued)

compd	V_{ss} (L/kg)	set	charge type	ref	compd	V_{ss} (L/kg)	set	charge type	ref
terbutaline sulfate	1.8	training	base	2	isosorbide dinitrate	3.9	test	neutral	2
tocainide	3	training	base	2	labetalol	9.4	test	zwitterion	2
tolmetin sodium	0.54	training	acid	2	levonorgestrel	1.7	test	neutral	2
trazodone	1	training	base	2	lincomycin	1.3	test	base	2
triamterene	13.4	training	neutral	2	melphalan	0.45	test	zwitterion	2
triazolam**	1.1	training	neutral	2	mercaptopurine	0.56	test	neutral	2
venlafaxine hydrochloride	7.5	training	base	2	methotrexate sodium	0.55	test	acid	2
verapamil	5	training	base	2	metoclopramide	3.4	test	base	2
zidovudine	1.4	training	neutral	2	metoprolol	4.2	test	base	2
zolpidem tartrate	0.54	training	neutral	2	misoprostol	14	test	neutral	2
amlodipine besylate	16	test	base	2	ondansetron	1.9	test	base	2
amphotericin b	0.76	test	zwitterion	2	pravastatin	0.04	test	acid	2
atenolol	0.95	test	base	2	prednisone	0.97	test	neutral	2
benazepril	0.12	test	zwitterion	2	propafenone	3.6	test	base	2
bumetanide	0.13	test	acid	2	quinidine gluconate	2.7	test	base	2
cefpodoxime proxetil	0.46	test	neutral	2	sertraline	76	test	base	2
cephradine	0.46	test	zwitterion	2	sulfisoxazole	0.15	test	acid	2
chlordiazepoxide	0.3	test	base	2	sulindac	2	test	acid	2
chloroquine	115	test	base	2	tetracycline	1.5	test	zwitterion	2
chlorpromazine	21	test	base	2	theophylline	0.5	test	neutral	2
clonazepam	3.2	test	neutral	2	timolol	2.1	test	base	2
cloxacillin	0.094	test	acid	2	tolbutamide	0.1	test	acid	2
clozapine	5.4	test	base	2	trimethoprim	1.6	test	base	2
dapsone	1	test	neutral	2	valproic acid	0.22	test	acid	2
desipramine	20	test	base	2	warfarin sodium	0.14	test	neutral	2
diclofenac	0.17	test	acid	2	zalcitabine	0.53	test	neutral	2
digitoxin	0.54	test	neutral	2					

[a] Ref 1: *J. Med. Chem.* **2004** *47*, 1242−1250; ref 2: *The Pharmacological Basis of Therapeutics*, 9th Ed.; Goodman & Gilman: 1996. **Indicates molecule excluded from model building due to descriptor failures.

Table 2. Human V_{ss} Training Set Statistics of the BNN, CART, and PLS Models[a]

model	r_0^2	r^2 (q^2)	rmse	ME
PLS	0.641	0.641 (0.597)	0.422	0.000
CART	0.871	0.876	0.253	0.016
BNN	0.790	0.794	0.323	0.018

[a] $N = 144$, standard deviation (SD) $y_{obs} = 0.71$, mean $y_{obs} = 0.124$. The r_0^2 is the coefficient of determination (correlation to line of unity), r^2/q^2 is the correlation to the line of best fit/cross validated r^2, rmse is the root mean square error, and ME is the mean error.

Table 3. Human V_{ss} Test Set Statistics for the BNN, CART, and PLS Models[a]

model	r_0^2	r^2	rmse	ME
PLS	0.577	0.587	0.494	0.032
CART	0.570	0.573	0.498	−0.028
BNN	0.550	0.560	0.509	0.075
CART−BNN−PLS	0.602	0.612	0.479	0.028

[a] $N = 50$, SD $y_{obs} = 0.77$, mean $y_{obs} = 0.133$. Also shown is the consensus prediction calculated as the average of the three individual model predictions.

the CART model. The r^2 value for the CART test set is larger compared to that of the PLS; however their r_0^2 and rmse values are similar indicating they have similar predictive power.

As has been demonstrated with the human derived model, our ability to predict experimental properties can often be improved using a consensus of different predictive algorithms. The differences in rmse between the best consensus model and the BNN model are smaller than the differences between the consensus model and the human model, suggesting the benefits of a consensus prediction in this case are less significant (Table 5 and Figure 4). It should be noted that in both cases the use of consensus predictions leads either to comparable or to better predictions than any single model but not to worse.

If we compare the results of the human and rat models solely using the r^2 or r_0^2 values, we would wrongly assume the model built on human data ($r^2 \sim 0.6$) is more accurate in prediction than the rat model ($r^2 \sim 0.5$). On the contrary, examining the rmse in prediction for the rat generated model, we see that the value is just ∼0.4 log units compared to the human model with an rmse of ∼0.5. This highlights that comparing r^2 or r_0^2 values between different models can be misleading. The r^2 and r_0^2 values depend on the variance of the data in each set, while the rmse value is absolute, allowing comparison of performance across different models and test sets.

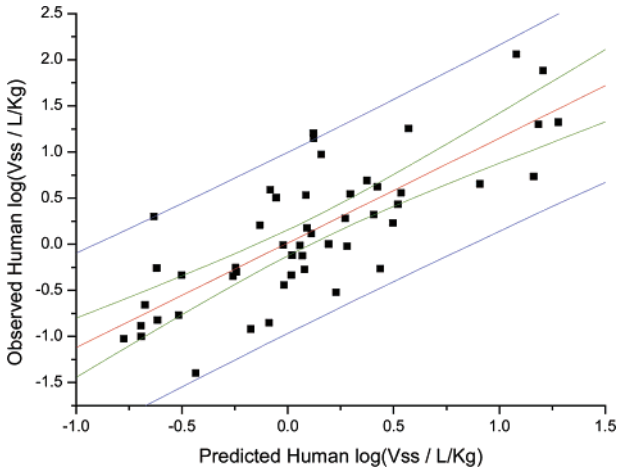

Figure 3. Plot of observed human log(V_{ss} (L/kg)) versus the in silico consensus prediction (CART-BNN-PLS) ($N = 50$). The regression line is colored red, with the 95% confidence limits given in green. The blue lines are the 95% confidence limits for the predictions. The regression line slope is 1.13, and the intercept is −0.02.

Table 4. Rat V_{ss} Training Set Statistics of the BNN, CART, and PLS Models[a]

model	r_0^2	r^2 (q^2)	rmse	ME
PLS	0.519	0.519 (0.506)	0.375	−0.001
CART	0.854	0.846	0.218	−0.01
BNN	0.767	0.767	0.261	−0.016

[a] $N = 1670$, SD $y_{obs} = 0.541$, mean $y_{obs} = 0.326$.

Table 5. Rat V_{ss} Test Set Statistics for the BNN, CART, and PLS Models[a]

model	r_0^2	r^2	rmse	ME
PLS	0.458	0.463	0.404	−0.037
CART	0.457	0.470	0.404	−0.027
BNN	0.519	0.527	0.380	−0.029
CART−BNN−PLS	0.534	0.538	0.374	−0.033

[a] N = 416, SD y_{obs} = 0.549, mean y_{obs} = 0.313. Also shown is the consensus prediction involving the average of all three models.

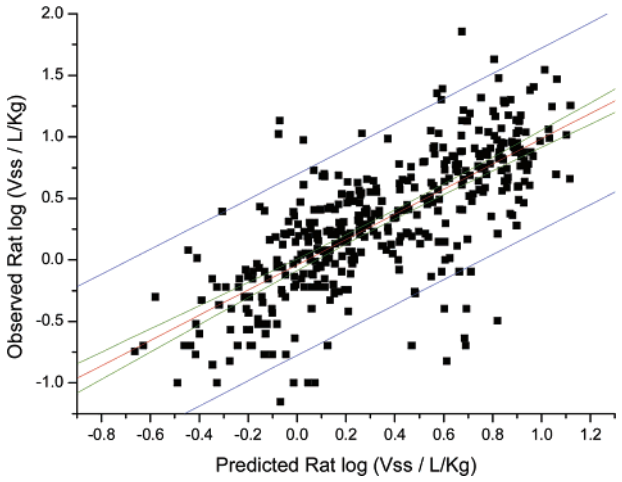

Figure 4. Plot of observed rat log(V_{ss} (L/kg)) versus the in silico consensus prediction (CART-BNN-PLS) (N = 416). The regression line is colored red, with the 95% confidence limits given in green. The blue lines are the 95% confidence limits for the predictions. Regression line slope is 1.03, and the intercept is −0.04.

Table 6. V_{ss} Ranges Often Assumed for Different Charge Types

	log(V_{ss} (L/kg))		
charge type	min	mean	max
acids (V_{ss}: 0.1−0.3 L/kg)	−1.00	−0.76	−0.52
bases (V_{ss}: >3 to >15 L/kg)	0.48	0.85	1.18
neutrals (V_{ss}: 0.5−2 L/kg)	−0.30	0.00	0.30
zwitterions (V_{ss}: 0.5−2 L/kg)	−0.30	0.00	0.30

3. V_{ss} Prediction from Charge State Alone. In the early stages of a project it is often assumed that the V_{ss} of a particular compound will lie within certain ranges based solely on charge state at pH 7.4 (Table 6). Acids for example are expected to lie within a relatively narrow V_{ss} range of between 0.1 and 0.3 L/kg in contrast to bases, which are initially assumed to have steady-state volumes greater than 3 L/kg. To ascertain whether a simple V_{ss} classification based on charge type alone was any better than either the in silico rat model or human model, we calculated the rmse in prediction using the mean value from the "expected ranges" as the best guess for the whole test set and each charge type individually (Table 7 and Table 8). We wished to assess whether our in silico models were an improvement on an in cerebro model, on the basis of an ADMET scientist's best estimate for volume of distribution of acids, neutrals, bases, and zwitterions as our prediction for each charge type. One could suggest that this is a truer test of predictive ability of a QSAR model than any other validation exercise.

3.1. Human Model. Analysis of the in silico model results for the test set in terms of ionization classes (Table 7) shows that zwitterions are predicted with the smallest error (0.43), followed by neutral (0.48) and basic compounds (0.48), with acids predicted with the largest error (0.50). Using the mean of the expected experimental V_{ss} ranges for the different charged species as the estimate of V_{ss}, we find that the rmse in prediction for the test set is 0.57 compared to 0.48 for the in silico model indicating the benefits of the in silico model.

Table 7. Human in Silico rmse and the rmse from a Prediction Based on Charge Type Only Are Reported for the Test Set[a]

human V_{ss} predictions	acid	base	neutral	zwitterion
N	9	19	15	7
in silico rmse	0.498	0.484	0.484	0.434
rmse (charge based prediction)	0.521	0.633	0.490	0.379
% of observations found in ranges	44	32	47	43

[a] The latter prediction is taken as the mean of the expected V_{ss} ranges as given in Table 6. Also shown are the percentages of observations that fall within the expected V_{ss} ranges for each charge class.

Table 8. Rat in Silico rmse and the rmse from a Prediction Based on Charge Type Only Are Reported for the Test Set[a]

rat V_{ss} predictions	acid	base	neutral	zwitterion
N	30	190	170	26
in silico rmse	0.344	0.392	0.359	0.375
rmse (charge based prediction)	0.505	0.499	0.437	0.565
% of observations found in ranges	31	72	59	42

[a] The latter prediction is taken as the mean of the expected V_{ss} ranges as given in Table 6. Also shown are the percentages of observations that fall within the expected V_{ss} ranges for each charge class.

When the analysis is done in terms of the individual charge types, we find that in all but one case the in silico model does better than that based solely on ionization state. The simplistic charge method predicts the seven zwitterions with an rmse of 0.38 compared to 0.43 for the in silico model. This may be a chance effect due to the small number of observations.

We also calculated the percentage of observations that lie within the expected V_{ss} ranges in a further effort to quantify the accuracy of the ranges. When we analyze the number of observations that fall within the expected V_{ss} ranges, we find that in the worst class, only 32% of bases fall within the expected ranges, while in the best predicted class, 47% of neutrals are correctly classified at best. This suggests the rather tight distributions expected for the different ionization states are unrealistic, and the model is doing significantly better than a scientist's intuitive guess (in cerebro prediction) just on the basis of charge-type.

3.2. Rat Model. Analysis of the in silico model results for the test set in terms of ionization classes (Table 8) shows a different trend to that observed from the human-based volume model. Acids are predicted with the smallest error (0.34), followed by neutral (0.36) and zwitterionic compounds (0.38), with bases predicted with the largest error (0.39). Using the mean of the expected experimental V_{ss} ranges for the different charged species as the estimate of V_{ss}, we find that the rmse in prediction for the test set is 0.48 compared to 0.37 for the in silico model indicating the benefits of the in silico method. Furthermore, when the analysis is done in terms of the individual ionization states, we find that in all cases the in silico model does better than that based solely on the mean value assumed for each charge type. When we analyze the number of observations that fall within the expected V_{ss} ranges for the particular ionization types, we find that at worst only 31% of acids fall within the ranges to 72% of bases at best.

4. Model Performance by Project. Project codes were extracted along with the in-house rat volume data to allow us to monitor the performance of the model for individual project

Table 9. Performance of the in Silico Rat V_{ss} Model by Project[a]

		mean log(V_{ss} (L/kg))			rmse	
project	*N*	expt	charge based prediction	in silico prediction	prediction based on charge type	in silico prediction
A	60	0.38	0.19	0.22	0.35	0.36
B	50	0.55	0.42	0.41	0.46	0.38
C	40	0.31	0.11	0.14	0.34	0.26
D	37	0.36	0.14	0.16	0.43	0.33
E	28	0.48	0.37	0.28	0.57	0.37
F	27	0.40	0.31	0.28	0.40	0.37
G	22	0.43	0.30	0.31	0.43	0.39
H	19	0.49	0.32	0.31	0.65	0.47
I	13	0.50	0.43	0.35	0.39	0.24
J	10	0.30	0.49	0.24	0.52	0.26

[a] The in silico model is also compared to a prediction based on charge type alone. Only projects where more than 10 observations are present are shown.

series. This is an important factor to consider as a global model often performs considerably better on some individual projects, and worse on others, as the overall test set statistics are averaged across all compounds and all project series. This can be both in absolute prediction and in terms of the rank order. We therefore computed the rmse in prediction of the test set compounds where a particular project had greater than 10 observations in the test set. We also report the mean experimental log(V_{ss}) observed for the project series, the mean predicted value from the in silico rat model, and the prediction based on the charge type only (Table 9).

We find projects C, J, and I all have rmses in prediction less than 0.26 log unit, which is considerably less than the global rat consensus model error would suggest at 0.37 log unit. In contrast, projects G and H have rmses larger than the global model error, suggesting it is desirable, where possible, to validate the QSAR model for a particular project series before using the in silico model in a project environment. In terms of predictions from ionization state alone, we find that apart from project A where the respective rmses are 0.36 and 0.35 log units, all projects are better predicted by the in silico model.

It is important to note that project compounds that are well represented in the test set (i.e., $N > 10$) will also be well represented in the training set due to the random training/test set selection, so the improved model performance must also be due in part to their better representation in the derived model. However, for projects where $N < 10$, we find many projects do better than the global model error, suggesting the model is not simply project specific.

5. Descriptors Controlling Steady-State Volume. 5.1. Human Model. There is a strong overlap between the types of descriptors identified using the three different statistical methodologies: BNN, CART, and PLS. To simplify matters, we only report the more easily interpretable PLS descriptors in the form of the scores (Figure 5), weights plot (Figure 6), and the overall normalized/scaled coefficients (Figure 7). The first component in the human model describes 51% of the variance in the training set and the second 13%. Component 1 essentially describes the lipophilicity/molecular weight dependency of V_{ss}, and the second component describes the charge state or charge distribution. Basic moieties/large positive charge and lipophilicity based descriptors have positive coefficients. This means that compounds with these features are expected to have higher volumes. In contrast, acidic/negative charge based descriptors have negative coefficients that generally lead to a decrease in volume. It must be noted that some descriptors are indicator variables (i.e., NEGCHARGED) while others have absolute negative values (i.e., average negative charge using Gasteiger−Huckel partial charge equilibration, Aver_Neg_Charge_G_H), so even though they are found to have different coefficient signs, they are actually having the same effect on the V_{ss}. Similarly, positive charge indicator variables have positive coefficients while the variance in positive charge over the van der Waals (VDW) surface area and the most positively charged atom in the molecule are negatively correlated with V_{ss}. The most positively charged atom in the molecule (MM_MAXPOS) correlates with the extent of delocalization (MM_VDW_EP_P_VAR), suggesting basic molecules may have varying V_{ss} driven by the extent of delocalization of the charge over the molecular surface.

5.2. Rat Model. In accord with the human V_{ss} model we find that the same descriptors are present in all three QSPR models and that charge type and lipophilicity dominate. We again focus on the PLS descriptors rather than the CART or BNN examples as these are more easily interpretable. Component 1 (45% of the explained variance) of the PLS model relates to lipophility/molecular weight, the second component (5% of the explained variance) to charge state/charge distribution, and the third component (2% of variance) to descriptors related to aromatic molecular features (Figure 8).

We find basic/positive charge and lipophilicity based descriptors have a positive coefficient which means compounds with these features are expected to have higher volumes. In contrast, acidic/negative charge descriptors have negative coefficients so generally lead to a decrease in volume. It must be noted that some descriptors are indicator variables (i.e., NEGCHARGED) and others have absolute negative values (i.e., average negative charge using Gasteiger−Marsili partial charge equilibration, Aver_Neg_Charge_Gast), so while they are found to have different coefficient signs, they are actually having the same effect on the volume.

6. Relationship between Human and Rat in Silico Models. A fundamental assumption in pharmacokinetics is that the unbound volume of distribution across species is constant. Therefore, volume of distribution differences across species are likely to be due to species differences in protein binding. The human and rat models were developed using different data sets compiled from two separate species, so the overlap between the models is an important aspect to consider, especially as the descriptors employed in both are so similar. On the basis of the analysis of 548 in house compounds with protein binding measurements in both species, it can be shown that while binding to plasma in both species is highly correlated ($r^2 = 0.78$), human protein binding is generally higher than in rat (Figure 9, Table 10), so in principle one could expect a systematic difference in the predictions from the two methods.

Predicting the rat test set of 416 compounds using the human model and vice versa shows the rat model predicts the human test set as well as the human model, both with an rmse of ~0.50 (Figures 10 and 11). However, the human model only predicts the rat test set with an rmse of ~0.50 compared to an rmse of ~0.38 from the rat model itself. The same result is obtained when either model is used to predict the other combined training/test set. Interestingly, the rat model predicts the human test set with no bias (low mean error), suggesting that the small protein binding differences between rat and human that may contribute to systematic bias in predicted V_{ss} are not identified by the model. To understand why the rat model predicts the human test set as well as the human model itself, we used 21 key descriptors, commonly used at AstraZeneca for comparative

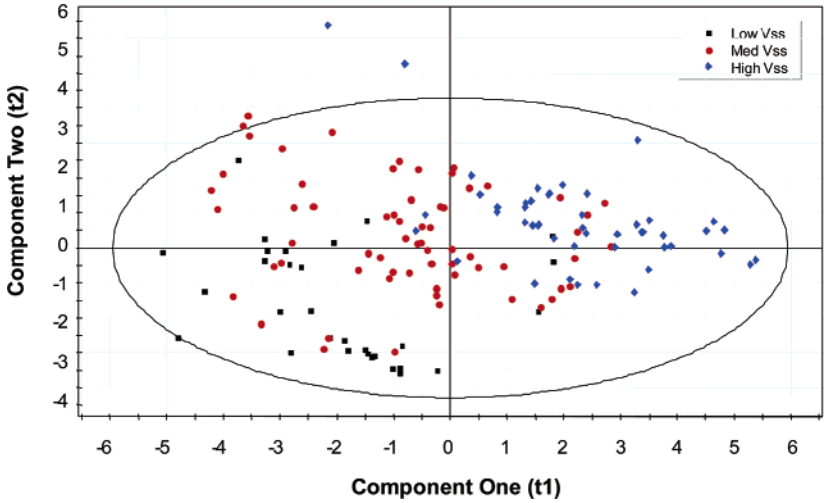

Figure 5. Human model PLS scores plot for the three-component PLS model. Only components 1 and 2 are shown as these account for the majority of the variance explained.

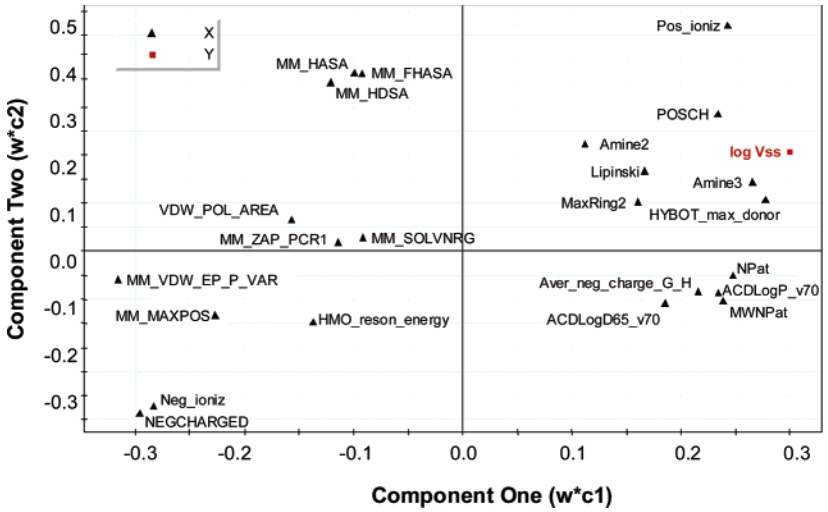

Figure 6. Human model PLS weights plot. Component 1 essentially encodes lipophilicity/molecular weight/size, and component 2 encodes charge type/hydrogen-bonding potential.

analysis, to assess the similarity between the two data sets. A PCA model fitting just two components describes 61% of the variance in the combined data set and clearly illustrates the differences between the data sets. From the PCA scores plot (Figure 12) it can be seen that while the majority of human compounds have rat near-neighbors, the converse is not true; there are many rat V_{ss} compounds that do not have human V_{ss} near-neighbors. This may explain why the rat model predicts the human results as well as the human model but not vice versa. Analysis of the PCA scores and loadings plots (Figures 12 and 13) shows that this difference is simply a reflection of the lower log *D*, log *P*, molecular weight (mol wt), calculated molar refractivity (CMR), and so forth of the literature derived compounds.

Conclusions

The human V_{ss} model explains ~50% of the variance of the test set of 50 compounds and has an error in prediction of ~0.47 log unit. The rat V_{ss} consensus model explains ~50% of the variance of the test set of 416 compounds and has an error in prediction of ~0.37 log unit. Both models perform better than a prediction of V_{ss} based solely on charge type and therefore could be utilized to estimate V_{ss} in either species in the early stages of a project.

The results herein suggest great care should be taken when using QSAR models developed on literature data of marketed drugs in the drug discovery process as these compounds are not generally representative of compounds commonly found in drug discovery research.

This is the first reported design and application of entirely in silico models for the prediction of an in vivo pharmacokinetic parameter. This approach demonstrates the potential of QSAR techniques, together with suitable high-quality data sets, to produce predictive ADME models, which may prove useful in the early stages of drug discovery prior to resource-intensive chemical synthesis and data acquisition.

Experimental Section

Computational Details. For this study PLS analyses were conducted using SIMCA 8.0[11] and GOLPE[12] and regression trees built using CART 4.0.[13] The CART methodology employs binary recursive partitioning, and in this model a consensus of 15 regression trees was found to be optimal when combined using bootstrapping aggregation (Bagging). The Gini algorithm together with least absolute deviation regression was used throughout this work. No misclassification costs were used in this analysis, and priors were set as equal. BNN models are less susceptible to overtraining and overfitting compared to classical neural networks.[14] Several publications have used BNN techniques for building QSAR

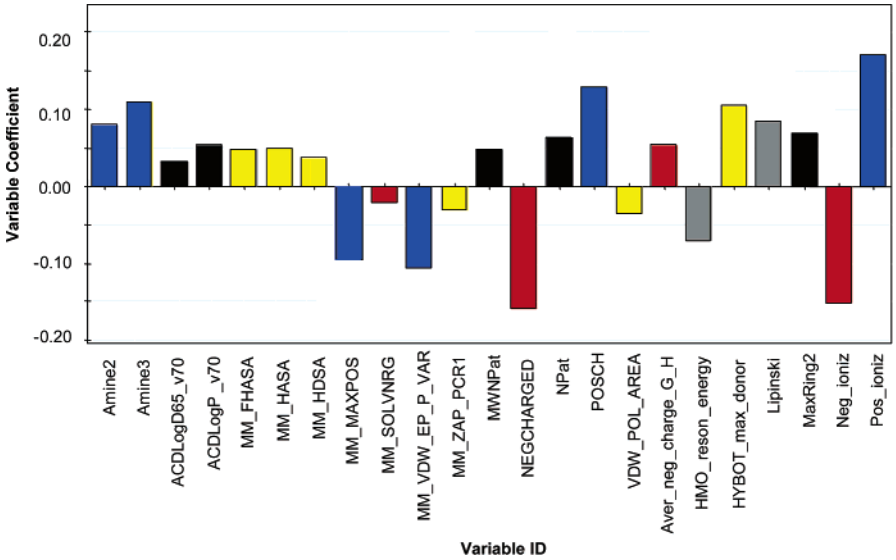

Figure 7. Coefficients derived from the descriptor PLS human V_{ss} model. ($r^2 = 0.64$, $q^2 = 0.60$, comp = 2). A broad classification of the descriptors is used above; red descriptors are acid/negative charge descriptors/indicators, blue descriptors are base/positive charge descriptors/indicators, black are lipophilicity/size based descriptors, gray are size/aromaticity based descriptors, and yellow are others.

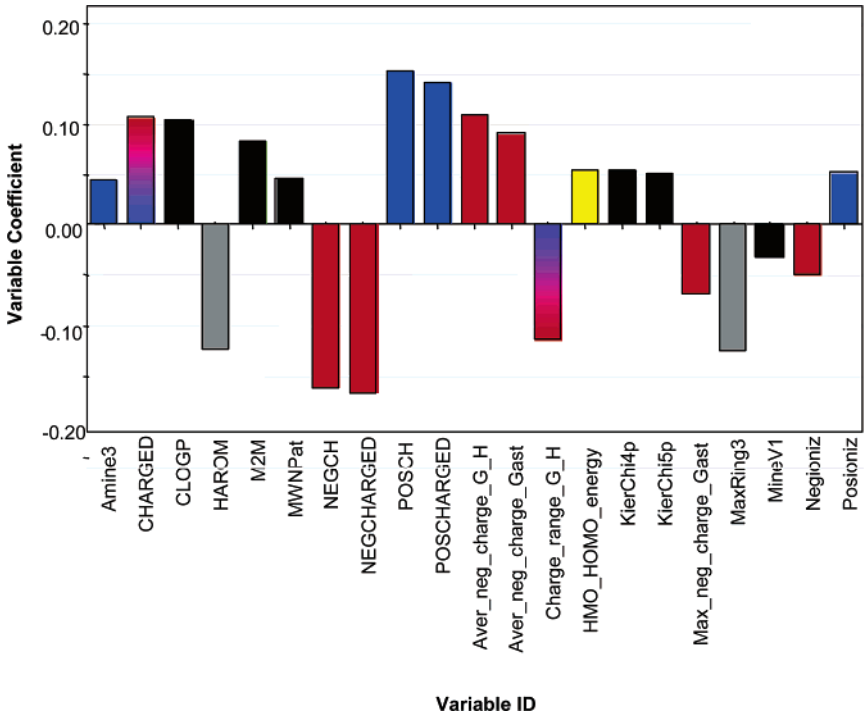

Figure 8. Coefficients derived from the descriptor rat V_{ss} PLS model ($r^2 = 0.52$, $q^2 = 0.51$, comp = 3). Red descriptors are acid/negative charge descriptors/indicators, blue descriptors are base/positive charge descriptors/indicators, black are lipophilicity/size based descriptors, gray are size/aromaticity based descriptors, and yellow are others.

models, and these techniques have been especially successful when applied to ADMET modeling.[15−18] For example, Sorich et al. have shown the BNN approach to produce superior models as compared to linear techniques when applied to the mapping of phase II metabolism.[18] A BNN model was produced using scripts in Perl language written by P. Bruneau,[19] coupled with an automated routine for variable selection written by R. Neal.[20] Prior to feeding the data into the BNN, both the descriptor vectors and the dependent variable were scaled to give a mean equal to 0 and a standard deviation equal to 1. The protocol followed to give a BNN model has been described by Bruneau.[19] This paper should be consulted for a full discussion on training parameters.

Intravenous Rat Pharmacokinetic Studies. All animal studies were conducted under U.K. Home Office License according to appropriate national legislation. Male rats (Sprague-Dawley; 150−250 g) were supplied by Charles River Ltd. (Margate, U.K.). Rats were given free access to food and water. Single dose (typically 1 mg/kg) plasma pharmacokinetics of AstraZeneca R&D Charnwood Discovery compounds were studied in rats after injection through the tail vein. Typically, blood samples (∼300 μL) were taken from the tail vein (reverse side to iv administration) at 2, 4, 8, 15, 30, 60, 120, 180, 300, 420, and 720 min postdose. Blood was collected in EDTA tubes, and plasma was removed following centrifugation (5 min, 4 °C, 3000 rpm). Plasma samples were analyzed by LC−MS/MS in MRM mode. Standard calibration curves were constructed by analyzing a series of control rat plasma aliquots containing suitable internal standard and various concentrations (1−40000 ng/mL) of test compound. The concentration of compound in each unknown sample was determined by solving the linear calibration curve equation for each corresponding drug/

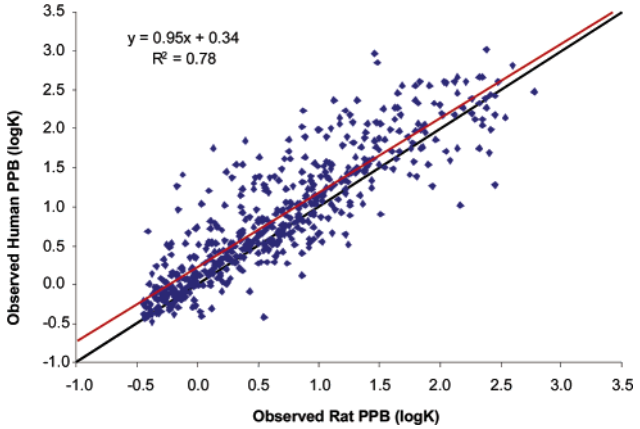

Figure 9. Plot of the observed human PPB against the experimental rat PPB for 548 compounds with measurements in both species. The line of unity is represented by the black line, with the line of best fit shown in red: rmse = 0.42 and ME = −0.19.

Table 10. Two Tailed *t* Test of Human and Rat Plasma Protein Binding (PPB) log *K* Measurements[a]

t test	human	rat
mean	0.558	0.527
variance	0.670	0.581
observations	548	
hypothesized mean difference	0.0	
P	9.16×10^{-5}	

[a] PPB results are generally higher in humans than in rat. Probability that they are significantly different is greater than the commonly used 99% confidence level.

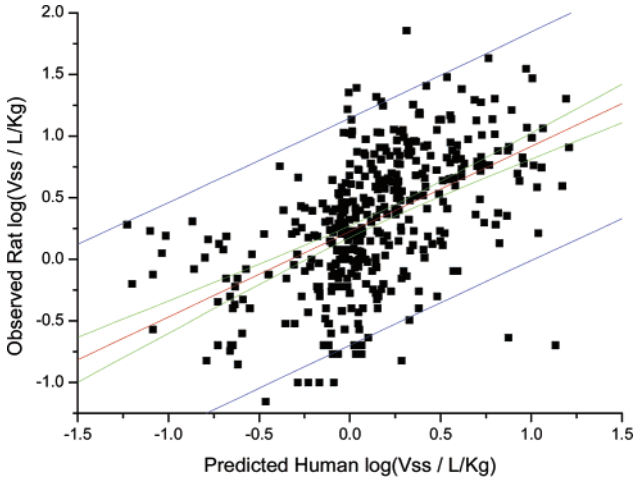

Figure 10. Human V_{ss} model predicting the rat V_{ss} test set ($N = 416$): $r^2 = 0.27$, $r_0^2 = 0.11$, rmse = 0.52, ME = −0.18, slope = 0.69, intercept = 0.22. Prediction of the total rat data set ($N = 2086$) gives $r^2 = 0.29$, $r_0^2 = 0.07$, rmse = 0.52, ME = −0.22, slope = 0.72, and intercept = 0.25.

internal standard ratio. Data manipulations and statistical calculations (mean ± SD) were performed in Excel software (Microsoft, WA). Plasma concentration versus time plots were analyzed using commercial software WinNonLin 3.1 (Pharsight, Mountain View, CA) to determine clearance, steady-state volume of distribution, and half-life.

Database Compilation. A total of 199 compounds of the human V_{ss} data set were obtained from two different literature sources[9,21] (Table 1). The compounds were coded as SMILES strings, and 123 descriptors, which broadly describe lipophilicity, size, topological, geometrical, and electronic features of molecules, were calculated using an AstraZeneca (AZ) molecular descriptor generator that has been described elsewhere.[22] Observations that reported errors on the key molecular descriptors such as ACDlogD7.4 were

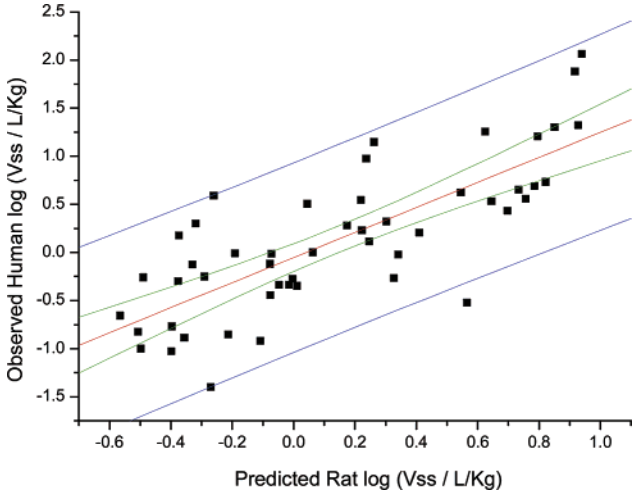

Figure 11. Rat V_{ss} model predicting the human V_{ss} test set ($N = 50$): $r^2 = 0.61$, $r_0^2 = 0.58$, rmse = 0.49, ME = −0.01, slope = 1.30, intercept = −0.06. Prediction of the total human data set ($N = 194$) gives $r^2 = 0.53$, $r_0^2 = 0.51$, rmse = 0.50, ME = −0.01, slope = 1.23, and intercept = −0.04.

removed as BNN and CART require complete matrices with no missing values. Furthermore, since we wished to compare the performance of the individual models against those of a consensus prediction, identical training and prediction sets were required. A data set of >2000 rat V_{ss} measurements was extracted from the internal databases at AstraZeneca R&D Charnwood. Compounds with nonquantitative V_{ss} measurements were excluded as we wished to build a continuous model, leaving 2086 compounds in total. Molecular descriptors were calculated and filtered using the same criteria used for the human data set.

The two data sets were randomly partitioned into training and test sets, 75%/25% for the human data and 80%/20% for rat. The V_{ss} values were logged to normalize the distribution of the errors across the V_{ss} range. The log(V_{ss} (L/kg)) distributions for the human and rat data sets are displayed in graphical form (Figures 1 and 2) for the four charge types: acid, base, neutral, and zwitterion.

QSAR Models. Human and rat PLS models were built and tested in SIMCA 8.0. Variable reduction was performed externally in the PLS software GOLPE by use of one round of *D*-optimal design (20% of variables removed on the basis of their coefficients) followed by fractional factorial selection, to find the key variables. Further variable removal was undertaken in SIMCA on the basis of an evaluation of the magnitude, the signs of the coefficients, and whether they agreed with chemical intuition. This led to a reduced-descriptor human model consisting of 2 components and 23 descriptors and a rat model with 3 components and 21 descriptors. The resulting models were the most significant PLS models in fit of the training set and the prediction of the test set.

To determine if the model could have occurred by chance, we performed randomization trials of the data matrix within SIMCA-P. The randomization tests were performed 500 times on the initial observed *y*-data, and the models were rebuilt. No randomized case approached the performance of our model in terms of r^2 or q^2 implying that our PLS models could not have occurred by chance.

Initial human and rat CART models were built using all the available descriptors and a consensus of 10 trees without pruning. The resulting models were assessed in fit and prediction of the training and test sets and subsequently refined in two ways: (a) using the average variable importance (VIP) of a descriptor from all the trees used in the consensus prediction and (b) using the optimal PLS descriptors. The CART VIP is a relative scale in which variables that are involved in primary splits have greater importance than those used further down the tree. We often find removal of variables with low importance can increase the predictivity of models. The optimal GOLPE derived descriptors can also be used

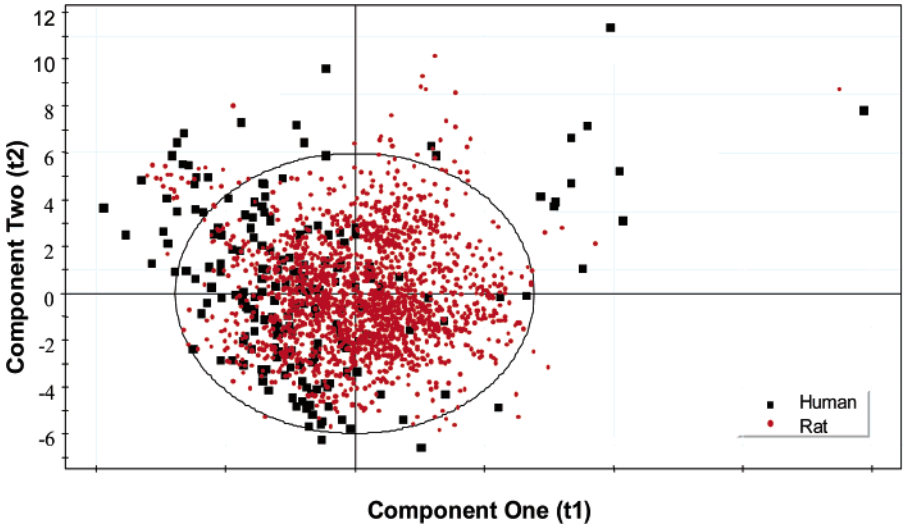

Figure 12. PCA model showing the relationship between the human and rat data sets using 23 key physicochemical descriptors. The first two components are shown, describing 61% of total variance in the data set: component 1 (35%) and component 2 (26%). The human data set differs significantly from the rat based data set on component 1.

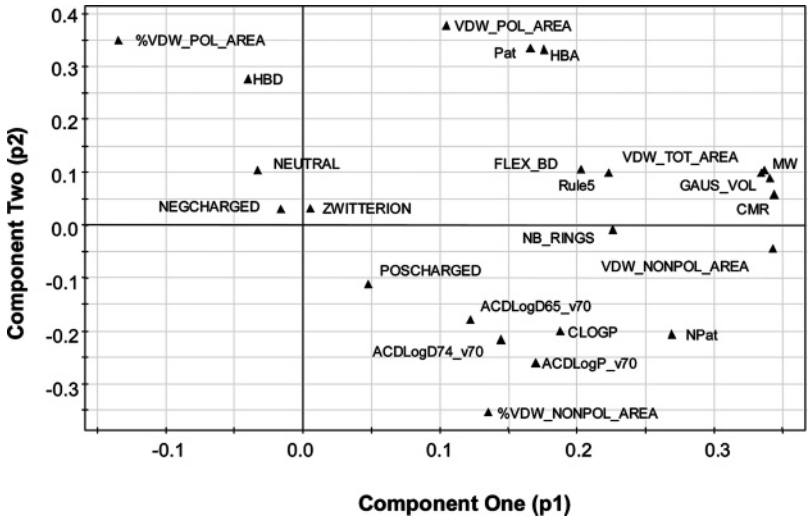

Figure 13. The PCA loadings plot. Component 1 encodes lipophilicity/molecular weight/size suggesting the rat data set (representing LO chemistry) is considerably higher in molecular weight/lipophilicity/rule-5-failures/hydrogen-bond acceptors than the human, predominantly marketed drugs data set.

as input for CART as these relate molecular structure to the response in a linear fashion. It is arguable whether using descriptors selected by a linear method is an acceptable input to a subsequent nonlinear modeling method. However, this is often applied pragmatically as a rapid variable reduction procedure, for instance, selecting variables by clustering or pairwise intercorrelation. Moreover, many nonlinear relationships can be approximated to some extent by a combination of linear models. Using the PLS descriptors as input for nonlinear methods often proves advantageous as a linear relationship is not imposed.

The optimal human model was obtained using the 30 descriptors with the highest VIP and a consensus of 10 separate trees. The optimal rat model obtained used the GOLPE derived descriptors (34) and a consensus of 10 separate trees.

Neural networks are computationally intensive methods that relate molecular characteristics to a response in either a linear or nonlinear fashion. The advantage of the Bayesian implementation of neutral nets above standard feed forward types is that a large number of networks are built using a distribution of starting weights and bias terms which are constantly updated during the modeling process. As the key descriptors and cross-terms are identified, the information is maintained within the model building process through the update of the starting distributions for the following cycle. Variable reduction is achieved using automatic relevance determination (ARD), which operates by removing descriptors whose weights contribute only a fraction of that compared to the most important example. A detailed description of the BNN implementation used at AZ is given elsewhere.[19,22]

The final models consist of a consensus of the last 200 cycles (models) of the training phase with the human model using 14 descriptors and the rat using 10 descriptors.

References

(1) Hodgson, J. ADMET—turning chemicals into drugs. *Nat. Biotechnol.* **2001**, *19* (8), 722−726.
(2) Lave, T.; Coassolo, P.; Reigner, B. Prediction of hepatic metabolic clearance based on interspecies allometric scaling techniques and in vitro−in vivo correlations. *Clin. Pharmacokinet.* **1999**, *36*, 211−231.
(3) Ward, K. W.; Smith, B. R. A comprehensive quantitative and qualitative evaluation of extrapolation of intravenous pharmacokinetic parameters from rat, dog and monkey to humans. II. Volume of distribution and mean residence time. *Drug Metab. Dispos.* **2004**, *32*, 612−619.
(4) Obach, R. S.; Baxter, J. G.; Liston, T. E.; Silber, B. M.; Jones, B. C.; MacIntyre, F.; Rance, D. J.; Wastall, P. The prediction of human pharmacokinetic parameters from preclinical and in vitro metabolism data. *J. Pharmacol. Exp. Ther.* **1997**, *283*, 46−58.

(5) Oie, S.; Tozer, T. N. Effect of altered plasma protein binding on apparent volume of distribution. *J. Pharm. Sci.* **1979**, *68* (9), 1203−1205.
(6) Austin, R. P.; Barton, P.; Mohmed, S.; Riley, R. J. The binding of drugs to hepatocytes and its relationship to physicochemical properties. *Drug Metab. Dispos.* **2005**, *33*, 419−425.
(7) Bjorkman, S. Prediction of the volume of distribution of a drug: Which tissue-plasma partition coefficients are needed? *J. Pharm. Pharmacol.* **2002**, *54*, 1237−1245.
(8) Wajima, T.; Fukumura, K.; Yano, Y.; Oguma, T. Prediction of human pharmacokinetics from animal data and molecular structural parameters using multivariate regression analysis: Volume of distribution at steady state. *J. Pharm. Pharmacol.* **2003**, *55*, 939−949.
(9) Lombardo, F.; Obach, R. S.; Shalaeva, M. Y.; Gao, F. Prediction of human volume of distribution values for neutral and basic drugs. 2. Extended data set and leave-class-out statistics. *J. Med. Chem.* **2004**, *47*, 1242−1250.
(10) Ghafourian, T.; Barzegar-Jalali, M.; Hakimiha, N.; Cronin, M. T. D. Quantitative structure-pharmacokinetic relationship modelling: Apparent volume of distribution. *J. Pharm. Pharmacol.* **2004**, *56*, 339−350.
(11) *SIMCA*, version 8.0; Umetrics AB (Tvistevägen 48, Box 7960 SE-907 19 Umeå, Sweden).
(12) Baroni, M.; Costatino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating optimal linear PLS estimations (GOLPE): an advanced chemometric tool for handling 3D-QSAR problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9−20.
(13) *CART*, version 4.0; Salford Systems (8880 Rio San Diego Dr., Suite 1045, San Diego, CA 92108).
(14) Sarle, W. ftp://ftp.sas.com/pub/neural/FAQ3.html.
(15) Ajay, W.; Murcko, M. Can We Learn To Distinguish Between "Drug-like" and "Nondrug-like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.
(16) Burden, F.; Winkler, D. Robust QSAR models using Bayesian regularized neural networks. *J. Med. Chem.* **1999**, *42*, 3183−3187.
(17) Burden, F.; Winkler, D. New QSAR methods applied to structure−activity mapping and combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 236−242.
(18) Sorich, M. J.; McKinnon, R. A.; Miners, J. O.; Winkler, D. A.; Smith, P. A. Rapid prediction of chemical metabolism by human UDP-glucuronosyltransferase isoforms using quantum chemical descriptors derived with the electronegativity equalization method. *J. Med. Chem.* **2004**, *47*, 5311−5317.
(19) Bruneau, P. Search for predictive generic model of aqueous solublity using Bayesian neural networks. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1605−1616.
(20) Neal, R. M. Software for Flexible Bayesian Modeling. http://www.cs.utoronto.ca/∼radford.
(21) *The Pharmacological Basis of Therapeutics*, 9th ed.; Goodman, L. S., Gilman, A. G., Eds.; McGraw-Hill Publishers: New York, 1996.
(22) Chohan, K.; Paine, S. W.; Mistry, J.; Barton, P.; Davis, A. M. A rapid computational filter for cytochrome P450 1A2 inhibition potential of compound libraries. *J. Med. Chem.* **2005**, *48*, 5154−5161.

JM0510070